

Layered Downlink Precoding for C-RAN Systems with Full Dimensional MIMO

Jinkyu Kang, Osvaldo Simeone, Joonhyuk Kang and Shlomo Shamai (Shitz)

Abstract

The implementation of a Cloud Radio Access Network (C-RAN) with Full Dimensional (FD)-MIMO is faced with the challenge of controlling the fronthaul overhead for the transmission of baseband signals as the number of horizontal and vertical antennas grows larger. This work proposes to leverage the special low-rank structure of FD-MIMO channel, which is characterized by a time-invariant elevation component and a time-varying azimuth component, by means of a layered precoding approach, so as to reduce the fronthaul overhead. According to this scheme, separate precoding matrices are applied for the azimuth and elevation channel components, with different rates of adaptation to the channel variations and correspondingly different impacts on the fronthaul capacity. Moreover, we consider two different Central Unit (CU) - Radio Unit (RU) functional splits at the physical layer, namely the conventional C-RAN implementation and an alternative one in which coding and precoding are performed at the RUs. Via numerical results, it is shown that the layered schemes significantly outperform conventional non-layered schemes, especially in the regime of low fronthaul capacity and large number of vertical antennas.

Index Terms

Cloud-Radio Access Networks (C-RAN), Full Dimensional (FD)-MIMO, fronthaul compression, layered precoding.

Jinkyu Kang and Joonhyuk Kang are with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST) Daejeon, South Korea (Email: kangjk@kaist.ac.kr and jhkang@ee.kaist.ac.kr).

O. Simeone is with the Center for Wireless Communications and Signal Processing Research (CWCSPP), ECE Department, New Jersey Institute of Technology (NJIT), Newark, NJ 07102, USA (Email: osvaldo.simeone@njit.edu).

S. Shamai (Shitz) is with the Department of Electrical Engineering, Technion, Haifa, 32000, Israel (Email: sshlomo@ee.technion.ac.il).

I. INTRODUCTION

The cloud radio access network (C-RAN) architecture consists of multiple radio units (RUs) connected via fronthaul links to a central unit (CU) that implements the protocol stack of the RUs, including baseband processing [1], [2]. C-RAN enables a significant reduction in capital and operating expenses, as well as an enhanced spectral efficiency by means of joint interference management at the physical layer across all connected RUs. Nevertheless, it is well recognized that the performance of this architecture is limited by the capacity and latency constraints of the fronthaul network connecting RUs and CU [1]–[4].

In a standard C-RAN implementation, the fronthaul links carry digitized baseband signals. Hence, the bit rate required for a fronthaul link is determined by the quantization and compression operations applied to the baseband signals prior to transmission on the fronthaul links. As such, the fronthaul rate is proportional to the signal bandwidth, to the oversampling factor, to the resolution of the quantizer/compressor, and to the number of antennas [5]. The fronthaul bit rate can be reduced by implementing alternative functional splits between CU and RU, whereby some baseband functionalities are implemented at the RU [6]–[8].

As a concurrent trend in the evolution of wireless networks, in the 3rd generation partnership project (3GPP) long term evolution (LTE) Release-13, three-dimensional (3D)-MIMO, where base stations are equipped with two-dimensional rectangular antenna arrays, has been intensely discussed as a promising tool to boost spectral efficiency [9], [10]. 3D-MIMO technology is classified into three categories, namely, vertical sectorization (VS), elevation beamforming (EB), and Full-Dimensional MIMO (FD-MIMO) in order of complexity. The VS scheme splits a sector of cellular coverage into multiple sectors by means of different electrical downtilt angles. With the EB approach, instead, users are supported by predetermined or adaptive beams in the elevation direction. Finally, in FD-MIMO, the spatial diversity provided by vertical and horizontal antennas is leveraged jointly to serve multiple users using multiuser-MIMO techniques.

Endowing RUs with two-dimensional arrays in a C-RAN system (see Fig. 1), while promising from a spectral efficiency perspective, creates significant challenges in terms of fronthaul overhead as the number of antennas grows larger [11]. In this paper, we focus on the design of downlink precoding for C-RANs

with FD-MIMO RUs by accounting for the impact of fronthaul capacity limitations. Previous works [4], [12]–[15] on precoding design for the downlink of C-RAN systems either assume fixed channel matrices with full channel state information (CSI), see [4], [12]–[14], or considers ergodic channels with generic correlation structure and possibly imperfect CSI [15]. Importantly, these works do not account for the special features of FD channel models [16], [17] and hence do not bring insights into the feasibility of a C-RAN deployment based on FD-MIMO. In particular, the FD-MIMO channel is understood to be characterized by time variability at different time scales for elevation and azimuth components; elevation component changes significantly more slowly than the rate of change of the more conventional azimuth component [16].

In order to address the design and performance of C-RAN system with FD-MIMO, this paper puts forth the following contributions.

- A *layered precoding* scheme is proposed whereby separate precoding matrices are applied for the azimuth and elevation channel components with a different rate of adaptation to the channel variations. Specifically, a single precoding matrix is designed for the elevation channel across all coherence times based on stochastic CSI, while precoding matrices are optimized for the azimuth channel by adapting instantaneous CSI. This layered approach, considered in [17] for a conventional cellular architecture, has the unique advantage in a C-RAN of potentially reducing the fronthaul transmission rate, due to the opportunity to amortize the overhead related to the elevation channel component across multiple coherence times.
- We study layered precoding in a C-RAN system by considering two different CU-RU functional splits at the physical layer, namely the conventional C-RAN implementation, referred to as Compress-After-Precoding (CAP) as in [4], [12]–[15], whereby all baseband processing is done at the CU, and an alternative split, known as Compress-Before-Precoding (CBP) [15], [18], in which channel encoding and precoding are instead performed at the RUs.
- We carry out a performance comparison between standard non-layered precoding strategies and

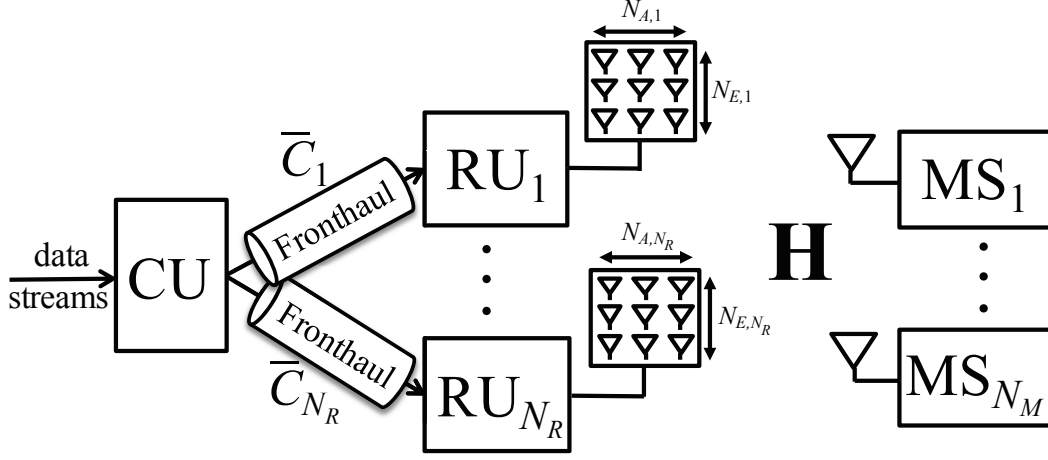


Fig. 1. Downlink of a C-RAN system with FD-MIMO.

layered precoding for C-RAN systems with FD-MIMO under different functional splits as a function of system parameters such as the fronthaul capacity and the duration of the coherence period.

The rest of the paper is organized as follows. We describe the system model in Section II. In Section III, we review the conventional non-layered precoding schemes corresponding to the mentioned functional splits, namely CAP and CBP [15]. Then, we propose and optimize the layered precoding strategy for fronthaul compression in Section IV. In Section V, numerical results are presented. Concluding remarks are summarized in Section VI.

Notation: $E[\cdot]$ and $\text{tr}(\cdot)$ denote the expectation and trace of the argument matrix, respectively. We use the standard notation for mutual information [19]. $\nu_{\max}(\mathbf{A})$ is the eigenvector corresponding to the largest eigenvalue of the semi-positive definite matrix \mathbf{A} . We reserve the superscript \mathbf{A}^T for the transpose of \mathbf{A} , \mathbf{A}^\dagger for the conjugate transpose of \mathbf{A} , and $\mathbf{A}^{-1} = (\mathbf{A}^\dagger \mathbf{A})^{-1} \mathbf{A}^\dagger$, which reduces to the usual inverse if the number of columns and rows are same. The identity matrix is denoted as \mathbf{I} . $\mathbf{A} \otimes \mathbf{B}$ is the Kronecker product of \mathbf{A} and \mathbf{B} .

II. SYSTEM MODEL

We consider the downlink of a C-RAN in which a cluster of N_R RUs provides wireless service to N_M mobile stations (MSs) as illustrated in Fig. 1. Each RU i has a FD, or two-dimensional (2D), antenna array of $N_{A,i}$ horizontal antennas by $N_{E,i}$ vertical antennas and each MS has a single antenna. RU i

is connected to the CU via fronthaul link of capacity \bar{C}_i bit per downlink symbol, where the downlink symbol rate equals the baud rate, i.e., no oversampling is performed.

A. Signal Model

Each coded transmission block spans multiple coherence periods, e.g., multiple distinct resource blocks in an LTE system, of the downlink channel that contain T symbols each. The $T \times 1$ signal \mathbf{y}_j received by the MS j in a given coherence interval is given by

$$\mathbf{y}_j = \mathbf{X}^T \mathbf{h}_j + \mathbf{z}_j, \quad (1)$$

where \mathbf{z}_j is the $T \times 1$ noise vector with i.i.d. $\mathcal{CN}(0, 1)$ components; $\mathbf{h}_j = [\mathbf{h}_{j1}^T, \dots, \mathbf{h}_{jN_R}^T]^T$ denotes the $\sum_{i=1}^{N_R} N_{A,i} N_{E,i} \times 1$ channel vector for MS j , where \mathbf{h}_{ji} is the $N_{A,i} N_{E,i} \times 1$ channel vector from the i -th RU to the MS j as further discussed below; and \mathbf{X} is an $\sum_{i=1}^{N_R} N_{A,i} N_{E,i} \times T$ matrix that stacks the signals transmitted by all the RUs, i.e., $\mathbf{X} = [\mathbf{X}_1^T, \dots, \mathbf{X}_{N_R}^T]^T$, where \mathbf{X}_i is a $N_{A,i} N_{E,i} \times T$ complex baseband signal matrix transmitted by the i -th RU with each channel coherence period of duration T channel uses. Note that each column of the signal matrix \mathbf{X}_i corresponds to the signal transmitted from the $N_{A,i} N_{E,i}$ antennas in a channel use. The transmit signal \mathbf{X}_i has a power constraint given as $E[|\mathbf{X}_i|^2] = T \bar{P}_i$.

The channel vector \mathbf{h}_j is assumed to be constant during each channel coherence block and to change according to a stationary ergodic process from block to block. We assume that the CU has perfect instantaneous information about the channel matrix $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_{N_M}]$ and MSs have full CSI about their respective channel matrices.

B. FD Channel Model

As in, e.g., [16], [17], we assume that each RU is equipped with a uniform rectangular array (URA). Furthermore, the channel vector \mathbf{h}_{ji} from RU i to MS j is modeled by means of a Kronecker product spatial correlation model [16], [17]. This was shown to provide a good modeling choice under the condition that the MS is sufficiently far away from the RUs [16]. According to this model, the covariance of the

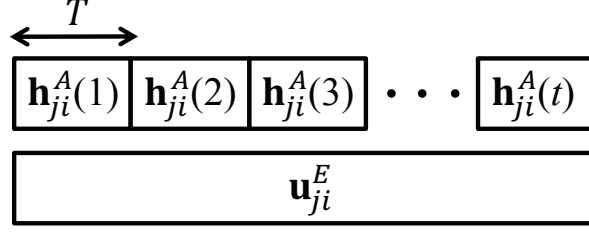


Fig. 2. Illustration of time variability of the azimuth component $\{\mathbf{h}_{ji}^A(t)\}$ and of the elevation component \mathbf{u}_{ji}^E in the FD channel model (3). The notation $\mathbf{h}_{ji}^A(t)$ emphasizes the dependence on the coherence block t of the azimuth component of the channel.

3D channel \mathbf{h}_{ji} which is defined as $\mathbf{R}_{ji} = E[\mathbf{h}_{ji}\mathbf{h}_{ji}^\dagger]$, is written as

$$\mathbf{R}_{ji} = \mathbf{R}_{ji}^A \otimes \mathbf{R}_{ji}^E, \quad (2)$$

where \mathbf{R}_{ji}^A and \mathbf{R}_{ji}^E represent the covariance matrices in the azimuth and elevation directions, respectively.

Since the elevation direction is typically subject to negligible scattering [20], [21], the elevation covariance matrix \mathbf{R}_{ji}^E may be assumed to be a rank-1 matrix, i.e., $\mathbf{R}_{ji}^E = \mathbf{u}_{ji}^E \mathbf{u}_{ji}^{E\dagger}$, where \mathbf{u}_{ji}^E is a $N_{E,i} \times 1$ unit-norm vector [17]. Under this assumption, the channel vector \mathbf{h}_{ji} can be written as

$$\mathbf{h}_{ji} = \sqrt{\alpha_{ji}} \mathbf{h}_{ji}^A \otimes \mathbf{u}_{ji}^E, \quad (3)$$

where α_{ji} denotes the path loss coefficient between MS j and RU i as

$$\alpha_{ji} = \frac{1}{1 + \left(\frac{d_{ji}}{d_0}\right)^\eta}, \quad (4)$$

with d_{ji} being the distance between the j -th MS and the i -th RU, d_0 being a reference distance, and η being the path loss exponent; and $\mathbf{h}_{ji}^A \sim \mathcal{CN}(0, \mathbf{R}_{ji}^A)$ with \mathbf{R}_{ji}^A having diagonal elements equal to one.

This model entails that the elevation components \mathbf{h}_{ji}^E remains constant over coherence interval, while the azimuth component changes independent across coherence interval as $\mathbf{h}_{ji}^A \sim \mathcal{CN}(0, \mathbf{R}_{ji}^A)$, as illustrated in Fig. 2.

III. BACKGROUND

In this section, we briefly recall in an informal fashion two baseline strategies for downlink transmission in the C-RAN system introduced above. The strategies correspond to two different functional splits at the physical layer between CU and RUs [5], [6] as detailed in [15]. We note that these schemes were

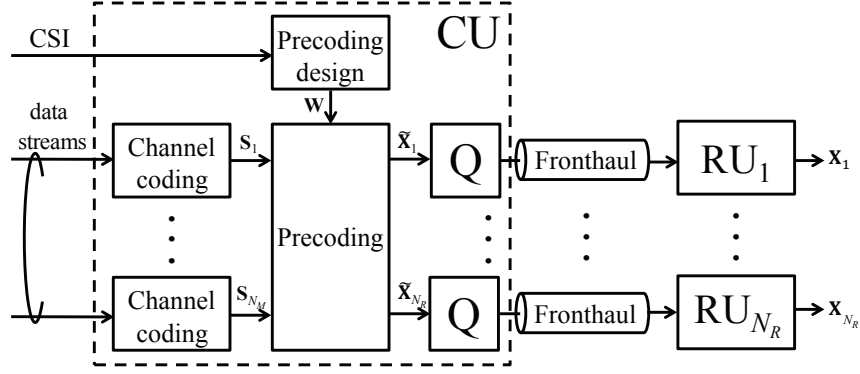


Fig. 3. Block diagram of the (non-layered) Compression-After-Precoding (CAP) scheme (“Q” represents fronthaul compression).

previously proposed and studied without specific reference to FD-MIMO and hence do not leverage the special structure of the channel model (3).

A. Standard C-RAN Processing: Precoding at the CU

In the standard C-RAN approach, all baseband processing is done at the CU. Specifically, as illustrated in Fig. 3, the CU performs channel coding and precoding, and then compresses the resulting baseband signals so that they can be forwarded on the fronthaul links to the corresponding RUs. The RUs upconvert the received quantized baseband signal prior to transmission on the wireless channel. Following [15], we refer to this strategy as Compression-After-Precoding (CAP). Analysis and optimization of the CAP strategy can be found in [15].

B. Alternative Functional Split: Precoding at the RUs

As an alternative to the standard C-RAN approach just described, one can instead implement channel encoding and precoding at the RUs. This is referred to as Compression-Before-Precoding (CBP) in [15], [18]. According to this solution, as seen in Fig. 4, the CU calculates the precoding matrices based on the available CSI, but does not perform precoding. Instead, it uses the fronthaul links to communicate the downlink information streams to each RU, along with the compressed precoding matrix. Each RU can then encode and precode the messages of the MSs based on the information received from the fronthaul link. As elaborated on in [15], this alternative functional split is generally advantageous when the number of MSs is not too large and when the coherence period T is large enough. This is because, when the number

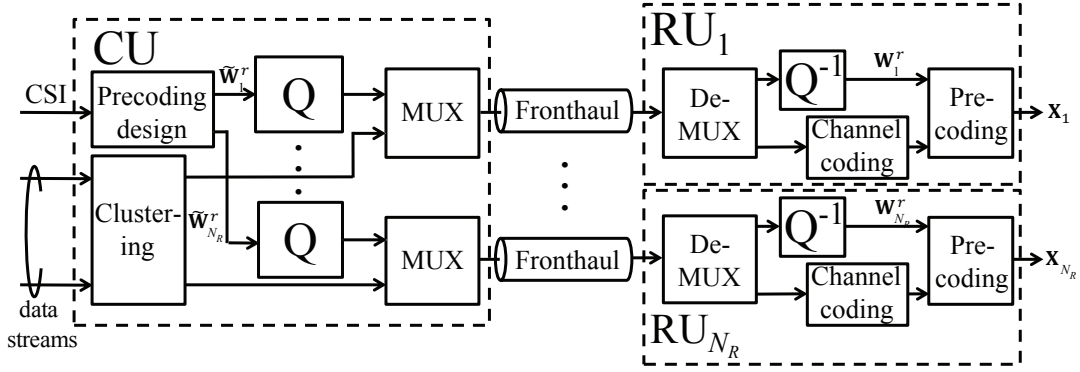


Fig. 4. Block diagram of the (non-layered) Compression-Before-Precoding (CBP) scheme (“Q” represents fronthaul compression).

of MSs is small, a lower fronthaul overhead is needed to communicate the data streams of the MSs on the fronthaul link; and, when the coherence period T is large, the compressed precoding information can be amortized over a longer period, hence reducing the fronthaul rate.

IV. LAYERED PRECODING FOR REDUCED FRONTHAUL OVERHEAD

The baseline state-of-the-art fronthaul transmission strategies mentioned above do not make any provision to exploit the special structure of the FD channel model (3), and can hence be inefficient if the number of vertical antennas is large. In this section, we propose a layered precoding that instead leverages the different dynamic characteristic of the elevation and azimuth channels as per channel model (3). We recall that, according to this model, the elevation channel has a constant direction across the coherence periods in its elevation component due to the rank-1 covariance matrix, while its azimuth component changes in each coherence period due to the generally larger rank of its covariance matrix (see Fig. 2).

In order to exploit this channel decomposition, we propose that the CU designs separate precoding matrices for the elevation and azimuth channels following a layered precoding approach. The key idea is that of designing a single precoding matrix for the elevation channel across all coherence times based on long-term CSI, while adapting only the azimuth precoding matrix to the instantaneous channel conditions. This allows the CU to accurately describe the elevation precoding matrix through the fronthaul links via quantization with negligible overhead given that the latter is amortized across all coherence periods. Precoding on the azimuth channel can instead be handled via either a CAP or CBP-like scheme, as detailed

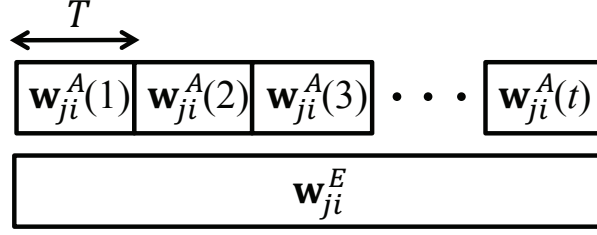


Fig. 5. Illustration of time variability of the azimuth and elevation components of beamforming in the layered precoding scheme (5).

below.

In the following, we first describe the layered precoding approach in Section IV-A; then introduce the precoding and fronthaul compression strategy based on CAP in Section IV-B; and, finally, we introduce CBP-based fronthaul compression and layered precoding design in Section IV-C.

A. Layered Precoding

Leveraging the channel decomposition resulting from the Kronecker channel model (3), we propose to factorize the $N_{t,i} \times 1$ precoding vector \mathbf{w}_{ji} for RU i toward MS j as

$$\mathbf{w}_{ji} = \mathbf{w}_{ji}^A \otimes \mathbf{w}_{ji}^E, \quad (5)$$

where \mathbf{w}_{ji}^A denotes the $N_{A,i} \times 1$ azimuth component and \mathbf{w}_{ji}^E is the $N_{E,i} \times 1$ elevation component of the precoding vector for MS j and RU i designed based on the elevation channels. A similar model was proposed in [17] for co-located antenna arrays. The corresponding $N_{A,i} \times N_M$ azimuth precoding matrix \mathbf{W}_i^A and the $N_{E,i} \times N_M$ elevation precoding matrix \mathbf{W}_i^E for RU i are defined as $\mathbf{W}_i^A = [\mathbf{w}_{1i}^A, \dots, \mathbf{w}_{N_M i}^A]$ and $\mathbf{W}_i^E = [\mathbf{w}_{1i}^E, \dots, \mathbf{w}_{N_M i}^E]$, respectively. In the proposed solutions, each elevation component \mathbf{w}_{ji}^E is quantized by the CU and sent to the j -th RU via the corresponding fronthaul links. Since this vector is to be used for all coherence times, as illustrated in Fig. 5, its fronthaul overhead can be amortized across multiple coherence interval. As a result, it can be assumed to be known accurately at the RUs. Moreover, the corresponding fronthaul overhead for the transfer of elevation precoding information on the fronthaul links can be assumed to be negligible. For the azimuth components, we may adopt either a CAP or CBP approach, as discussed next.

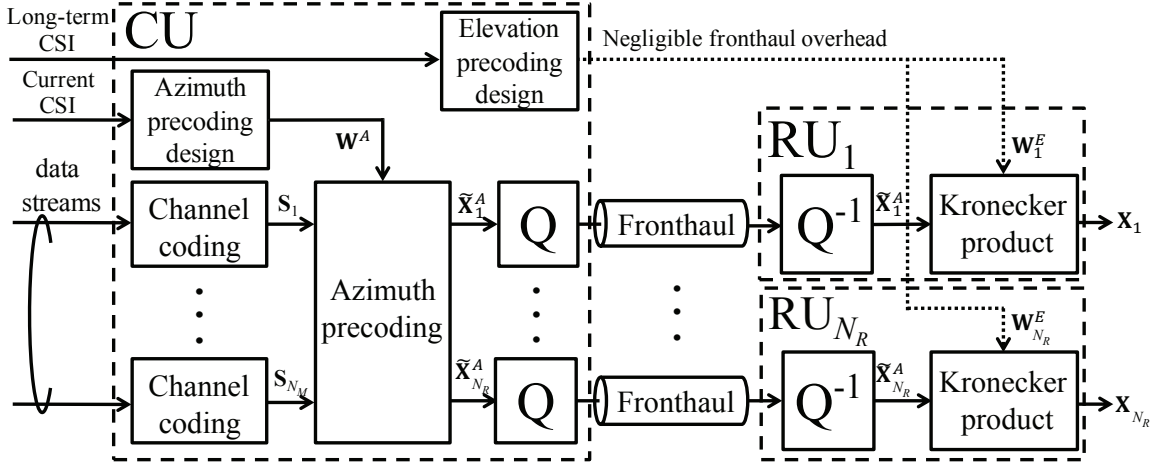


Fig. 6. Block diagram of the Layered Compression-After-Precoding (CAP) scheme (“Q” represents fronthaul compression).

B. CAP-based Fronthaul Compression for Layered Precoding

In the proposed CAP-based solution, the CU applies precoding only for the azimuth component. Accordingly, the azimuth-precoded baseband signals, as well as the precoding matrix for the elevation component, are separately compressed at the CU and forwarded over the fronthaul links to each RU. In order to perform precoding over both elevation and azimuth channels, each RU finally performs the Kronecker product of the compressed baseband signal \mathbf{X}_{ji}^A and the precoding vector \mathbf{w}_{ji}^E for elevation channel. A block diagram can be found in Fig. 6 and details are provided next.

1) *Details and Analysis:* Let $\tilde{\mathbf{X}}_{ji}^A$ be the $N_{A,i} \times T$ precoded signal only for the azimuth channel between RU i and MS j in a given coherence period. This is defined as $\tilde{\mathbf{X}}_{ji}^A = \mathbf{w}_{ji}^A \mathbf{s}_j^T$, where \mathbf{s}_j is the $T \times 1$ vector containing the encoded data stream for MS j in the given coherence period. Note that all the entries of vector \mathbf{s}_j are assumed to have i.i.d. $\mathcal{CN}(0, 1)$ from standard random coding arguments. Adopting a CAP-like approach, the CU quantizes each sequence of baseband signals $\{\tilde{\mathbf{X}}_{ji}^A\}$, for all $j \in \mathcal{N}_M$, across all coherence periods intended for RU i for transfer on i -th fronthaul. The compressed signal \mathbf{X}_{ji}^A is modeled as

$$\mathbf{X}_{ji}^A = \tilde{\mathbf{X}}_{ji}^A + \mathbf{Q}_{x,ji}^A = \mathbf{w}_{ji}^A \mathbf{s}_j^T + \mathbf{Q}_{x,ji}^A, \quad (6)$$

where $\mathbf{Q}_{x,ji}^A$ is the quantization noise matrix, which is assumed to have i.i.d. $\mathcal{CN}(0, \sigma_{x,ji}^2)$ entries. From standard rate-distortion arguments [19], [22], the required rate for transfer of the precoded data signals

$\{\tilde{\mathbf{X}}_{ji}^A\}_{j \in \mathcal{N}_M}$ on fronthaul link between the CU and RU i is given as

$$C_{x,i}(\mathbf{W}_i^A, \boldsymbol{\sigma}_{x,i}^2) = \sum_{j=1}^{N_M} I(\mathbf{X}_{ji}^A; \tilde{\mathbf{X}}_{ji}^A) = \sum_{j=1}^{N_M} \{\log(\|\mathbf{w}_{ji}^A\|^2 + \sigma_{x,ji}^2) - \log \sigma_{x,ji}^2\}, \quad (7)$$

where we have used the assumption that the data signal \mathbf{X}_{ji}^A are independent across the MS index j and we have defined $\boldsymbol{\sigma}_{x,i}^2 = [\sigma_{x,1i}^2, \dots, \sigma_{x,N_M i}^2]^T$. Note that, unlike the standard CAP scheme, here the signals for different MSs are separately compressed as per (6).

Considering also the elevation component, the resulting signal \mathbf{X}_i computed and transmitted by RU i is obtained as $\mathbf{X}_i = \sum_{j=1}^{N_M} \mathbf{X}_{ji}$, with

$$\mathbf{X}_{ji} = \mathbf{X}_{ji}^A \otimes \mathbf{w}_{ji}^E = (\mathbf{w}_{ji}^A \mathbf{s}_j^T + \mathbf{Q}_{x,ji}^A) \otimes \mathbf{w}_{ji}^E = (\mathbf{w}_{ji}^A \otimes \mathbf{w}_{ji}^E) \mathbf{s}_j^T + \mathbf{Q}_{x,ji}^A \otimes \mathbf{w}_{ji}^E. \quad (8)$$

The power transmitted at RU i is then computed as

$$\begin{aligned} P_i(\mathbf{W}_i^A, \mathbf{W}_i^E, \boldsymbol{\sigma}_{x,i}^2) &= \text{tr}(\mathbf{X}_i \mathbf{X}_i^\dagger) = \text{tr} \left(\sum_{j=1}^{N_M} ((\mathbf{w}_{ji}^A \mathbf{s}_j^T + \mathbf{Q}_{x,ji}^A) \otimes \mathbf{w}_{ji}^E) ((\mathbf{w}_{ji}^A \mathbf{s}_j^T + \mathbf{Q}_{x,ji}^A) \otimes \mathbf{w}_{ji}^E)^\dagger \right) \\ &= \sum_{j=1}^{N_M} (\|\mathbf{w}_{ji}^A\|^2 \|\mathbf{w}_{ji}^E\|^2 + N_{A,i} \sigma_{x,ji}^2 \|\mathbf{w}_{ji}^E\|^2), \end{aligned} \quad (9)$$

where we have used the property of the Kronecker product that $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC} \otimes \mathbf{BD})$ and $\text{tr}(\mathbf{A} \otimes \mathbf{B}) = \text{tr}(\mathbf{A})\text{tr}(\mathbf{B})$ [23].

The ergodic achievable rate for MS j is evaluated as $E[R_j(\mathbf{H}, \mathbf{W}^A, \mathbf{W}^E, \boldsymbol{\sigma}_x^2)]$, with $R_j(\mathbf{H}, \mathbf{W}^A, \mathbf{W}^E, \boldsymbol{\sigma}_x^2) = I_{\mathbf{H}}(\mathbf{s}_j; \mathbf{y}_j)/T$, where $I_{\mathbf{H}}(\mathbf{s}_j; \mathbf{y}_j)$ is the mutual information conditioned on the value of channel matrix \mathbf{H} , the expectation is taken with respect to \mathbf{H} and

$$\begin{aligned} R_j(\mathbf{H}, \mathbf{W}^A, \mathbf{W}^E, \boldsymbol{\sigma}_x^2) &= \log \left(1 + \sum_{k=1}^{N_M} \sum_{i=1}^{N_R} \lambda_{ji}^E |\mathbf{u}_{ji}^E \mathbf{w}_{ki}^E|^2 \left(|\mathbf{w}_{ki}^A \mathbf{h}_{ji}^A|^2 + \sigma_{x,ki}^2 \|\mathbf{h}_{ji}^A\|^2 \right) \right) \\ &\quad - \log \left(1 + \sum_{k=1, k \neq j}^{N_M} \sum_{i=1}^{N_R} \lambda_{ji}^E |\mathbf{u}_{ji}^E \mathbf{w}_{ki}^E|^2 \left(|\mathbf{w}_{ki}^A \mathbf{h}_{ji}^A|^2 + \sigma_{x,ki}^2 \|\mathbf{h}_{ji}^A\|^2 \right) \right), \end{aligned} \quad (10)$$

where $\mathbf{W}^A = [(\mathbf{W}_1^A)^T, \dots, (\mathbf{W}_{N_R}^A)^T]^T$, $\mathbf{W}^E = [(\mathbf{W}_1^E)^T, \dots, (\mathbf{W}_{N_R}^E)^T]^T$, and $\boldsymbol{\sigma}_x^2 = [\sigma_{x,1}^2, \dots, \sigma_{x,N_R}^2]$.

2) *Problem Formulation:* The ergodic achievable sum-rate (10) can be optimized over the precoding matrices \mathbf{W}^A and \mathbf{W}^E , and over the quantization noise variance vector $\boldsymbol{\sigma}_x^2$ under fronthaul capacity and power constraints. Since the design of the precoding matrix \mathbf{W}^A for azimuth channel and of the

Algorithm 1 CAP-based Fronthaul Compression and Layered Precoding Design

1) Long-term Optimization of Elevation Precoding

Input: Long-term statistics of the channel

Output: Elevation precoding \mathbf{W}^{E*}

Initialization (outer loop): Initialize the covariance matrix $\mathbf{V}^{E(n)} \succeq 0$ subject to $\text{tr}(\mathbf{V}^{E(n)}) = 1$ and set $n = 0$.

Repeat

$n \leftarrow n + 1$

Generate a channel matrix realization $\mathbf{H}^{(n)}$ using the available stochastic CSI.

Inner loop: Obtain $\mathbf{V}^{A(n)}(\mathbf{H}^{(n)})$ and $\boldsymbol{\sigma}_x^{2(n)}(\mathbf{H}^{(n)})$ with $\mathbf{V}^E \leftarrow \mathbf{V}^{E(n-1)}$ using Algorithm 2.

Update $\mathbf{V}^{E(n)}$ by solving problem (23), which depends on $\mathbf{V}^{A(m)}(\mathbf{H}^{(m)})$ and $\boldsymbol{\sigma}_x^{2(m)}(\mathbf{H}^{(m)})$

for all $m \leq n$.

Until a convergence criterion is satisfied.

Set $\mathbf{V}^E \leftarrow \mathbf{V}^{E(n)}$.

Calculation of \mathbf{W}^{E*} : Calculate the precoding matrix \mathbf{W}^{E*} for elevation channel from the covariance matrix \mathbf{V}^E via rank reduction as $\mathbf{w}_{ji}^{E*} = \nu_{\max}(\mathbf{V}_{ji}^E)$ for all $j \in \mathcal{N}_M$ and $i \in \mathcal{N}_R$.

2) Short-term Optimization of Azimuth Precoding and Quantization Noise

Input: Channel \mathbf{H} and elevation precoding \mathbf{W}^{E*}

Output: Azimuth precoding $\mathbf{W}^{A*}(\mathbf{H})$ and quantization noise vector $\boldsymbol{\sigma}_x^{2*}(\mathbf{H})$

Obtain $\mathbf{V}^A(\mathbf{H})$ and $\boldsymbol{\sigma}_x^2(\mathbf{H})$ with $\mathbf{W}^E \leftarrow \mathbf{W}^{E*}$ using Algorithm 2.

Calculation of $\mathbf{W}^{A*}(\mathbf{H})$: Calculate the precoding matrix $\mathbf{W}^{A*}(\mathbf{H})$ for the azimuth channel from the covariance matrix $\mathbf{V}^A(\mathbf{H})$ via rank reduction as $\mathbf{w}_{ji}^{A*}(\mathbf{H}) = \beta_{ji} \nu_{\max}(\mathbf{V}_{ji}^A(\mathbf{H}))$ for all $j \in \mathcal{N}_M$ and $i \in \mathcal{N}_R$, where β_{ji} is obtained by imposing $P_i(\mathbf{W}_i^{A*}(\mathbf{H}), \mathbf{W}_i^{E*}, \boldsymbol{\sigma}_{x,i}^{2*}(\mathbf{H})) = \bar{P}_i$ using (9).

compression noise variance $\boldsymbol{\sigma}_x^2$ is adapted to the channel realization \mathbf{H} for each coherence block, we use the notations $\mathbf{W}^A(\mathbf{H})$ and $\boldsymbol{\sigma}_x^2(\mathbf{H})$. The problem of maximizing the achievable rate is then formulated as follows

$$\begin{aligned} & \underset{\mathbf{W}^A(\mathbf{H}), \mathbf{W}^E, \boldsymbol{\sigma}_x^2(\mathbf{H})}{\text{maximize}} && \sum_{j \in \mathcal{N}_M} E[R_j(\mathbf{H}, \mathbf{W}^A(\mathbf{H}), \mathbf{W}^E, \boldsymbol{\sigma}_x^2(\mathbf{H}))] \end{aligned} \quad (11a)$$

$$\text{s.t.} \quad C_{x,i}(\mathbf{W}_i^A(\mathbf{H}), \boldsymbol{\sigma}_{x,i}^2(\mathbf{H})) \leq \bar{C}_i, \quad \forall i \in \mathcal{N}_R, \quad (11b)$$

$$P_i(\mathbf{W}_i^A(\mathbf{H}), \mathbf{W}_i^E, \boldsymbol{\sigma}_{x,i}^2(\mathbf{H})) \leq \bar{P}_i, \quad \forall i \in \mathcal{N}_R, \quad (11c)$$

Algorithm 2 DC Algorithm for Optimization of $\mathbf{V}^A(\mathbf{H})$ and $\boldsymbol{\sigma}_x^2(\mathbf{H})$

Input: Channel \mathbf{H} and elevation precoding \mathbf{V}^E .

Output: $\mathbf{V}^A(\mathbf{H})$ and $\boldsymbol{\sigma}_x^2(\mathbf{H})$

Initialization: Initialize $\mathbf{V}^{A(0)}(\mathbf{H}) \succeq 0$ and $\boldsymbol{\sigma}_x^{2(0)}(\mathbf{H}) \in \mathbb{R}^+$, and set $l = 0$.

Repeat

$l \leftarrow l + 1$

Update $\mathbf{V}^{A(l)}(\mathbf{H})$ and $\boldsymbol{\sigma}_x^{2(l)}(\mathbf{H})$ by solving problem (20).

Until a convergence criterion is satisfied.

Set $\mathbf{V}^A(\mathbf{H}) \leftarrow \mathbf{V}^{A(l)}(\mathbf{H})$ and $\boldsymbol{\sigma}_x^2(\mathbf{H}) \leftarrow \boldsymbol{\sigma}_x^{2(l)}(\mathbf{H})$.

where the constraints apply for all channel realizations \mathbf{H} , and we recall that the capacity constraint on i -th fronthaul link is \bar{C}_i and the power constraint for RU i is \bar{P}_i .

3) *Optimization Algorithm:* In problem (11), the objective function (11a) and constraint (11b) are non-convex in terms of $\mathbf{W}^A(\mathbf{H})$, \mathbf{W}^E , and $\boldsymbol{\sigma}_x^2(\mathbf{H})$. Furthermore, as discussed above, \mathbf{W}^E is designed based on stochastic CSI (long-term CSI), while $\mathbf{W}^A(\mathbf{H})$ and $\boldsymbol{\sigma}_x^2(\mathbf{H})$ are adapted to instantaneous CSI (short-term CSI). In order to tackle this problem, we propose an algorithm that optimizes separately the long-term and short-term variables \mathbf{W}^E and $(\mathbf{W}^A(\mathbf{H}), \boldsymbol{\sigma}_x^2(\mathbf{H}))$, respectively. For the former optimization, we adopt a stochastic optimization approach based empirical approximation of the ensemble averages in (11a) following Stochastic Successive Upper-bound Minimization (SSUM) method [24]. For the latter, we instead invoke the Difference of Convex (DC) method [25], [26] by leveraging the rank relaxation in obtained by reformulating the optimization problem in terms of the covariance matrices $\mathbf{V}_{ji}^A(\mathbf{H}) = \mathbf{w}_{ji}^A(\mathbf{H})\mathbf{w}_{ji}^{A\dagger}(\mathbf{H})$ and $\mathbf{V}_{ji}^E = \mathbf{w}_{ji}^E\mathbf{w}_{ji}^{E\dagger}$ for all $j \in \mathcal{N}_M$ and $i \in \mathcal{N}_R$. The resulting algorithm is detailed in Algorithm 1 and Appendix A. Note that, in Algorithm 1, long-term optimization has two nested loops in which inner loop requires at each iteration the solution of a convex problem, whose complexity is polynomial in the problem size [27].

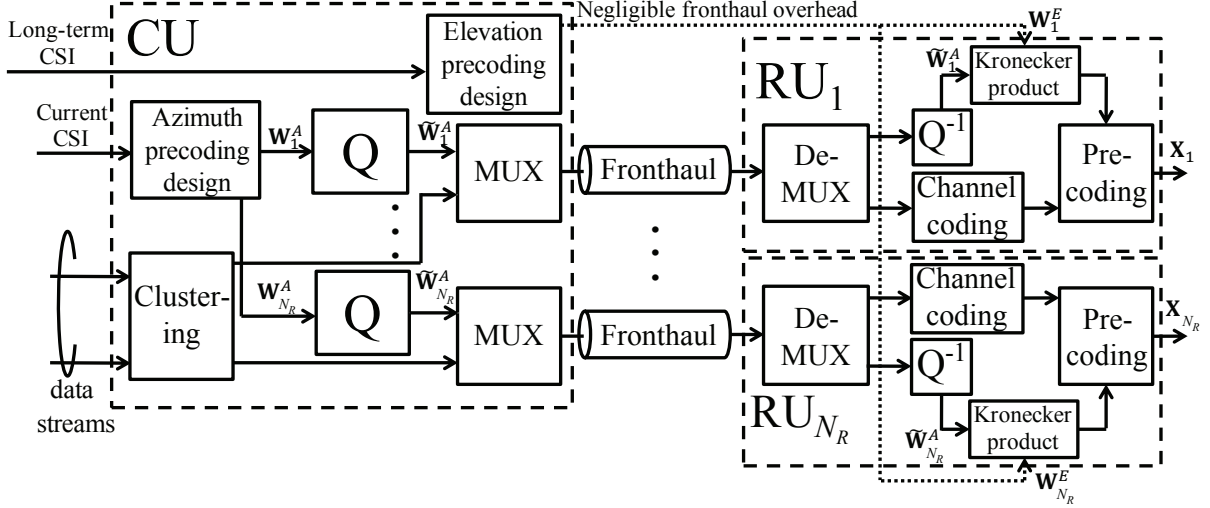


Fig. 7. Block diagram of the Layered Compression-Before-Precoding (CBP) scheme (“Q” represents fronthaul compression).

C. CBP-based Fronthaul Compression for Layered Precoding

In the proposed CBP-based strategy, as illustrated in Fig. 7, the CU designs the precoding matrices for both azimuth and elevation components, which are transferred, along with a given subset of downlink information messages, over the fronthaul link to the each RU. As discussed, since the design of the elevation precoding is done based on long-term CSI, and hence entails the use of a negligible portion of the fronthaul capacity, the fronthaul overhead depends only on the azimuth precoding matrices, which are adapted to current CSI, and on the information messages. As in [15], the subset of information messages sent to each RU is determined by a preliminary clustering step at the CU whereby each RU is assigned to serve a subset of the MSs. At each RU, the precoding matrix for FD-MIMO is computed via the Kronecker product between the precoding matrices for the azimuth and elevation channels. Based on the calculated precoding matrix, each RU can then encode and precode the received messages of the assigned MSs. Details are provided next.

1) *Details and Analysis:* To elaborate, let us denote the set of MSs assigned by RU i as $\mathcal{M}_i \subseteq \mathcal{N}_M$, for all $i \in \mathcal{N}_R$. We also use $\mathcal{M}_i[k]$ to denote the k -th MS in the set \mathcal{M}_i . Note that we assume that the assignment of MSs is given and not subject to optimization. The azimuth precoding vectors $\widetilde{\mathbf{W}}_i^A$ intended for RU i are compressed by the CU and forwarded over the fronthaul link to RU i . The compressed

azimuth precoding \mathbf{W}_i^A for RU i at the CU is then given by

$$\mathbf{W}_i^A = \widetilde{\mathbf{W}}_i^A + \mathbf{Q}_{w,i}, \quad (12)$$

where the quantization noise matrix $\mathbf{Q}_{w,i}$ is assumed to have zero-mean i.i.d. $\mathcal{CN}(0, \sigma_{w,i}^2)$ entries. The required rate for the transfer of the azimuth precoding on fronthaul link is given, similar to (7), as

$$\begin{aligned} C_{w,i}(\widetilde{\mathbf{W}}_i^A, \sigma_{w,i}^2) &= \frac{1}{T} I(\mathbf{W}_i^A; \widetilde{\mathbf{W}}_i^A) \\ &= \frac{1}{T} \{ \log \det (\widetilde{\mathbf{W}}_i^A \widetilde{\mathbf{W}}_i^{A\dagger} + \sigma_{w,i}^2 \mathbf{I}) - \log \det (\sigma_{w,i}^2 \mathbf{I}) \}, \end{aligned} \quad (13)$$

where $\widetilde{\mathbf{W}}_i^A = [\widetilde{\mathbf{w}}_{\mathcal{M}_i[1]i}^A, \dots, \widetilde{\mathbf{w}}_{\mathcal{M}_i[|\mathcal{M}_i|]i}^A]$. The remaining fronthaul capacity is used to convey information messages, whose total rate is $\sum_{j \in \mathcal{M}_i} R_j$ with R_j being the user rate for MS j . At each RU i , the precoding matrix for FD-MIMO is obtained via the Kronecker product of the elevation and azimuth components, yielding the transmitted signal $\mathbf{X}_i = \sum_{j \in \mathcal{M}_i} \mathbf{X}_{ji}$, with

$$\mathbf{X}_{ji} = (\mathbf{w}_{ji}^A \otimes \mathbf{w}_{ji}^E) \mathbf{s}_j^T = (\widetilde{\mathbf{w}}_{ji}^A \otimes \mathbf{w}_{ji}^E) \mathbf{s}_j^T + \mathbf{q}_{w,ji}^A \mathbf{s}_j^T \otimes \mathbf{w}_{ji}^E. \quad (14)$$

The power transmitted at RU i is then calculated as

$$P_i(\widetilde{\mathbf{W}}_i^A, \mathbf{W}_i^E, \sigma_{w,i}^2) = \text{tr}(\mathbf{X}_i \mathbf{X}_i^\dagger) = \sum_{j \in \mathcal{M}_i} (||\mathbf{w}_{ji}^A||^2 ||\mathbf{w}_{ji}^E||^2 + N_{A,i} \sigma_{w,i}^2 ||\mathbf{w}_{ji}^E||^2). \quad (15)$$

The ergodic achievable rate for MS j is calculated as $E[\bar{R}_j(\mathbf{H}, \widetilde{\mathbf{W}}^A, \mathbf{W}^E, \boldsymbol{\sigma}_w^2)]$ with

$$\begin{aligned} \bar{R}_j(\mathbf{H}, \widetilde{\mathbf{W}}^A, \mathbf{W}^E, \boldsymbol{\sigma}_w^2) &= \log \left(1 + \sum_{i=1}^{N_R} \sum_{k \in \mathcal{M}_i} \lambda_{ji}^E |\mathbf{u}_{ji}^E \mathbf{w}_{ki}^E|^2 \left(|\widetilde{\mathbf{w}}_{ki}^{A\dagger} \mathbf{h}_{ji}^A|^2 + \sigma_{w,i}^2 ||\mathbf{h}_{ji}^A||^2 \right) \right) \\ &\quad - \log \left(1 + \sum_{i=1}^{N_R} \sum_{k \in \mathcal{M}_i \setminus j} \lambda_{ji}^E |\mathbf{u}_{ji}^E \mathbf{w}_{ki}^E|^2 \left(|\widetilde{\mathbf{w}}_{ki}^{A\dagger} \mathbf{h}_{ji}^A|^2 + \sigma_{w,i}^2 ||\mathbf{h}_{ji}^A||^2 \right) \right), \end{aligned} \quad (16)$$

where $\widetilde{\mathbf{W}}^A = [\widetilde{\mathbf{W}}_1^{AT}, \dots, \widetilde{\mathbf{W}}_{N_R}^{AT}]^T$ and $\boldsymbol{\sigma}_w^2 = [\sigma_{w,1}^2, \dots, \sigma_{w,N_R}^2]$.

2) *Problem Formulation:* As discussed in Section IV-B, the azimuth precoding $\widetilde{\mathbf{W}}^A(\mathbf{H})$ and the compression noise variance $\sigma_w^2(\mathbf{H})$ can be adapted to the current channel realization at each coherence block. Accordingly, the optimization problem of interest can be formulated as

Algorithm 3 CBP-based Fronthaul Compression and Layered Precoding Design

1) Long-term Optimization of Elevation Precoding and User Rates

Input: Long-term statistics of the channel and clustering $\{\mathcal{M}_i\}$

Output: Elevation precoding \mathbf{W}^{E*} and MSs' rates $\{R_j\}$

Initialization (outer loop): Initialize the covariance matrix $\mathbf{V}^{E(n)} \succeq 0$ subject to $\text{tr}(\mathbf{V}^{E(n)}) = 1$ and $\{R_j^{(n)}\} \in \mathbb{R}^+$, and set $n = 0$.

Repeat

$n \leftarrow n + 1$

Generate a channel matrix realization $\mathbf{H}^{(n)}$ using the available stochastic CSI.

Inner loop: Obtain $\tilde{\mathbf{V}}^{A(n)}(\mathbf{H}^{(n)})$ and $\sigma_w^{2(n)}(\mathbf{H}^{(n)})$ with $\mathbf{V}^E \leftarrow \mathbf{V}^{E(n-1)}$ using Algorithm 4.

Update $\mathbf{V}^{E(n)}$ and $\{R_j^{(n)}\}$ by solving problem (29), which depends on $\tilde{\mathbf{V}}^{A(m)}(\mathbf{H}^{(m)})$ and $\sigma_w^{2(m)}(\mathbf{H}^{(m)})$ for all $m \leq n$.

Until a convergence criterion is satisfied.

Set $\mathbf{V}^E \leftarrow \mathbf{V}^{E(n)}$ and $\{R_j\} \leftarrow \{R_j^{(n)}\}$.

Calculation of \mathbf{W}^{E*} : Calculate the precoding matrix \mathbf{W}^{E*} for elevation channel from the covariance matrix \mathbf{V}^E via rank reduction as $\mathbf{w}_{ji}^{E*} = \nu_{\max}(\mathbf{V}_{ji}^E)$ for all $j \in \mathcal{N}_M$ and $i \in \mathcal{N}_R$.

2) Short-term Optimization of Azimuth Precoding and Quantization Noise

Input: Channel \mathbf{H} and elevation precoding \mathbf{W}^{E*}

Output: Azimuth precoding $\tilde{\mathbf{W}}^{A*}(\mathbf{H})$ and quantization noise vector $\sigma_w^{2*}(\mathbf{H})$

Obtain $\tilde{\mathbf{V}}^A(\mathbf{H})$ and $\sigma_w^2(\mathbf{H})$ with $\mathbf{W}^E \leftarrow \mathbf{W}^{E*}$ using Algorithm 4.

Calculation of $\tilde{\mathbf{W}}^{A*}(\mathbf{H})$: Calculate the precoding matrix $\tilde{\mathbf{W}}^{A*}(\mathbf{H})$ for the azimuth channel from the covariance matrix $\tilde{\mathbf{V}}^A(\mathbf{H})$ via rank reduction as $\tilde{\mathbf{w}}_{ji}^{A*}(\mathbf{H}) = \beta_{ji} \nu_{\max}(\tilde{\mathbf{V}}_{ji}^A(\mathbf{H}))$ for all $j \in \mathcal{N}_M$ and $i \in \mathcal{N}_R$, where β_{ji} is obtained by imposing $P_i(\tilde{\mathbf{W}}_i^{A*}(\mathbf{H}), \mathbf{W}_i^E, \sigma_{w,i}^{2*}(\mathbf{H})) = \bar{P}_i$ using (15).

$$\begin{aligned} & \underset{\tilde{\mathbf{W}}^A(\mathbf{H}), \mathbf{W}^E, \{R_j\}, \sigma_w^2(\mathbf{H})}{\text{maximize}} && \sum_{j \in \mathcal{N}_M} R_j \end{aligned} \quad (17a)$$

$$\text{s.t.} \quad R_j \leq E[\bar{R}_j(\mathbf{H}, \tilde{\mathbf{W}}^A(\mathbf{H}), \mathbf{W}^E, \sigma_w^2(\mathbf{H}))], \quad \forall j \in \mathcal{N}_M, \quad (17b)$$

$$C_{w,i}(\tilde{\mathbf{W}}_i^A(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H})) \leq \bar{C}_i - \sum_{j \in \mathcal{M}_i} R_j, \quad \forall i \in \mathcal{N}_R, \quad (17c)$$

$$P_i(\tilde{\mathbf{W}}_i^A(\mathbf{H}), \mathbf{W}_i^E, \sigma_{w,i}^2(\mathbf{H})) \leq \bar{P}_i, \quad \forall i \in \mathcal{N}_R, \quad (17d)$$

where the constraints apply to every channel realization \mathbf{H} .

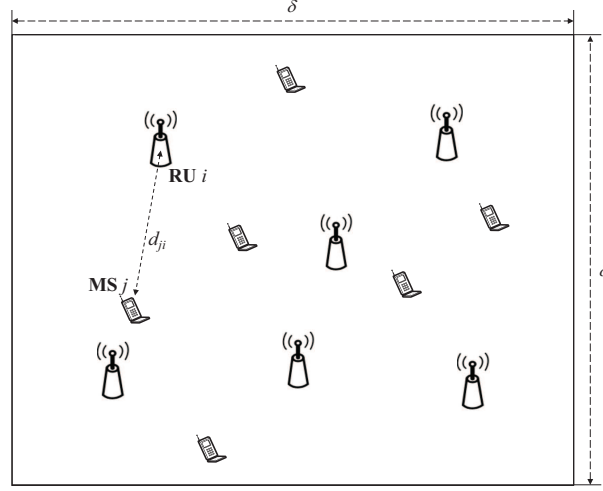


Fig. 8. Simulation environment for the numerical results.

Algorithm 4 DC Algorithm for Optimization of $\tilde{\mathbf{V}}^A(\mathbf{H})$ and $\sigma_w^2(\mathbf{H})$

Input: Channel \mathbf{H} and elevation precoding \mathbf{V}^E .

Output: $\tilde{\mathbf{V}}^A(\mathbf{H})$ and $\sigma_w^2(\mathbf{H})$

Initialization: Initialize $\tilde{\mathbf{V}}^{A(0)}(\mathbf{H}) \succeq 0$ and $\sigma_w^{2(0)}(\mathbf{H}) \in \mathbb{R}^+$, and set $l = 0$.

Repeat

$l \leftarrow l + 1$

Update $\tilde{\mathbf{V}}^{A(l)}(\mathbf{H})$ and $\sigma_w^{2(l)}(\mathbf{H})$ by solving problem (26).

Until a convergence criterion is satisfied.

Set $\tilde{\mathbf{V}}^A(\mathbf{H}) \leftarrow \tilde{\mathbf{V}}^{A(l)}(\mathbf{H})$ and $\sigma_w^2(\mathbf{H}) \leftarrow \sigma_w^{2(l)}(\mathbf{H})$.

3) *Optimization Algorithm:* Similar to Section IV-B, the non-convex functions $\bar{R}_j(\mathbf{H}, \tilde{\mathbf{W}}^A(\mathbf{H}), \mathbf{W}^E, \sigma_w^2(\mathbf{H}))$ and $C_{w,i}(\tilde{\mathbf{W}}^A(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H}))$ can be seen to be DC functions of the covariance matrices $\tilde{\mathbf{V}}_{ji}^A(\mathbf{H}) = \tilde{\mathbf{w}}_{ji}^A(\mathbf{H})\tilde{\mathbf{w}}_{ji}^{A\dagger}(\mathbf{H})$ and $\mathbf{V}_{ji}^E = \mathbf{w}_{ji}^E\mathbf{w}_{ji}^{E\dagger}$ for all $j \in \mathcal{N}_M$ and $i \in \mathcal{N}_R$. Moreover, the optimization problem can be divided into long-term and short-term optimizations, that can be tackled via the SSUM and DC methods, respectively, as summarized in Algorithm 3 and detailed in Appendix B. Moreover, as in Algorithm 1, it is required to solve one convex problem, which has polynomial complexity [27], at each inner iteration.

V. NUMERICAL RESULTS

In this section, we compare the performance of the strategies with layered precoding, namely layered CAP and CBP schemes, and the conventional strategies, namely CAP and CBP schemes, for FD-MIMO

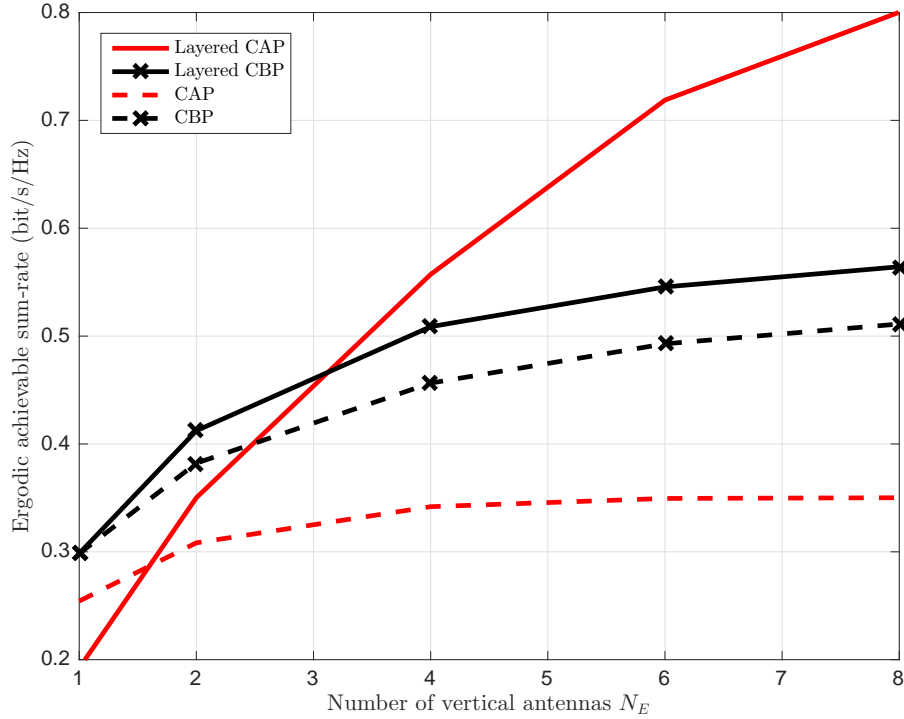


Fig. 9. Ergodic achievable sum-rate vs. the number of vertical antennas N_E ($N_R = N_M = 2$, $N_{A,i} = 2$, $C = 1$ bit/s/Hz, $P = 0$ dB, and $T = 20$).

systems. To this end, we consider a set-up simulation environment where the RUs and MSs are randomly located in a square area with side $\delta = 500$ m as in Fig. 8. In the path loss formula (4), we set the reference distance to $d_0 = 50$ m and the path loss exponent to $\eta = 3$ with d_{ji} being the Euclidean distance between the i -th RU and the j -th MS. The channels are assumed to have the Kronecker model in (3). Throughout, we assume that the every RU is subject to the same fronthaul capacity \bar{C} and has the same power constraint \bar{P} , namely $\bar{C}_i = \bar{C}$ and $\bar{P}_i = \bar{P}$ for $i \in \mathcal{N}_R$. Throughout, we consider CBP strategies in which each RU serves all MSs, i.e., $N_C = N_M$.

Fig. 9 shows the ergodic achievable sum-rate as function of the number of vertical antennas N_E , where the number of RUs and MSs is $N_R = N_M = 2$, the number of horizontal antennas is $N_{A,i} = 2$ for all $i \in \mathcal{N}_R$, the fronthaul capacity is $\bar{C} = 1$ bit/s/Hz, the transmit power is $\bar{P} = 0$ dB and the coherence time is $T = 20$. We observe that the layered precoding schemes provide increasingly large gains as N_E grows larger. This is because, in the conventional strategies, the fronthaul overhead for the transfer of elevation

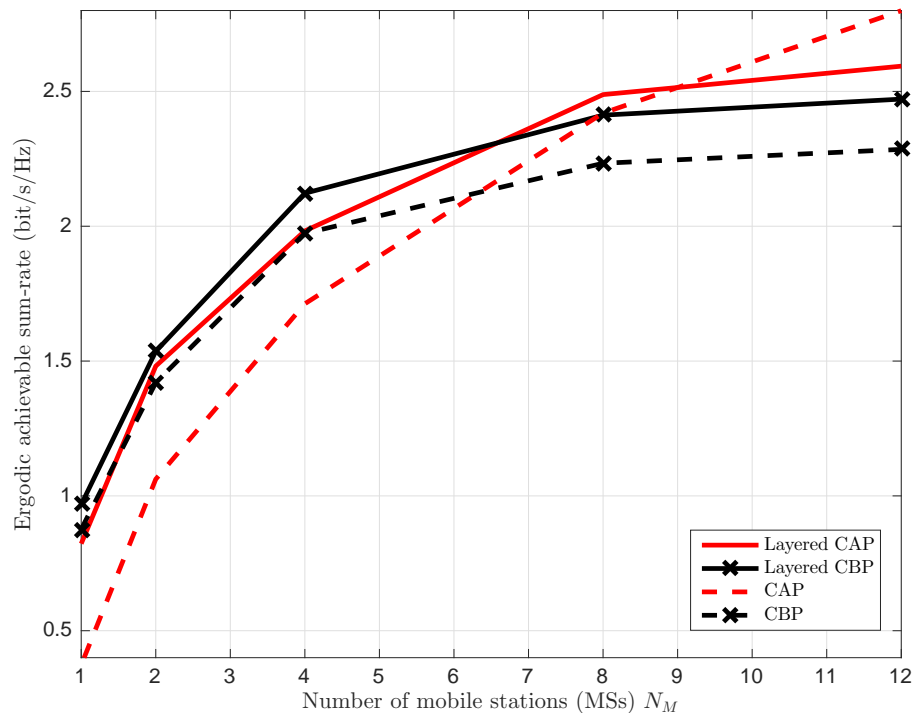


Fig. 10. Ergodic achievable sum-rate vs. the number of MSs N_M ($N_R = 2$, $N_{A,i} = 2$, $N_{E,i} = 4$, $C = 3$ bit/s/Hz, $P = 5$ dB, and $T = 20$).

precoding information increases with the number of vertical antennas. This gain is less pronounced here for layered CBP strategies, whose achievable rate is limited here by the relatively small coherence interval, as further discussed below (see also Sec. III-B). Moreover, it is observed that, for $N_E = 1$, the conventional and the layered precoding strategies with CBP method have the same performance, while this is not the case for the CAP strategies. In fact, the conventional CAP strategy outperforms the layered CAP strategy for small values of N_E . This is caused by the fact that, with the layered CAP strategy, the azimuth precoded signals for the MSs are separately compressed, hence entailing an inefficient use of the fronthaul when N_E is large enough.

Fig. 10 shows the effect of the number of MSs N_M on the ergodic achievable sum-rate with $N_R = 2$, $N_{A,i} = 2$ and $N_{E,i} = 4$ for all $i \in \mathcal{N}_R$, $C = 3$ bit/s/Hz, $P = 5$ dB, and $T = 20$. The CBP methods show the known poor performance as the number of MSs increases, due to the need for the transmission of the messages of all MSs on all fronthaul links [15]. Moreover, in keeping with the discussion above, we observe that the conventional CAP method is to be preferred in the regime of large number of MSs.

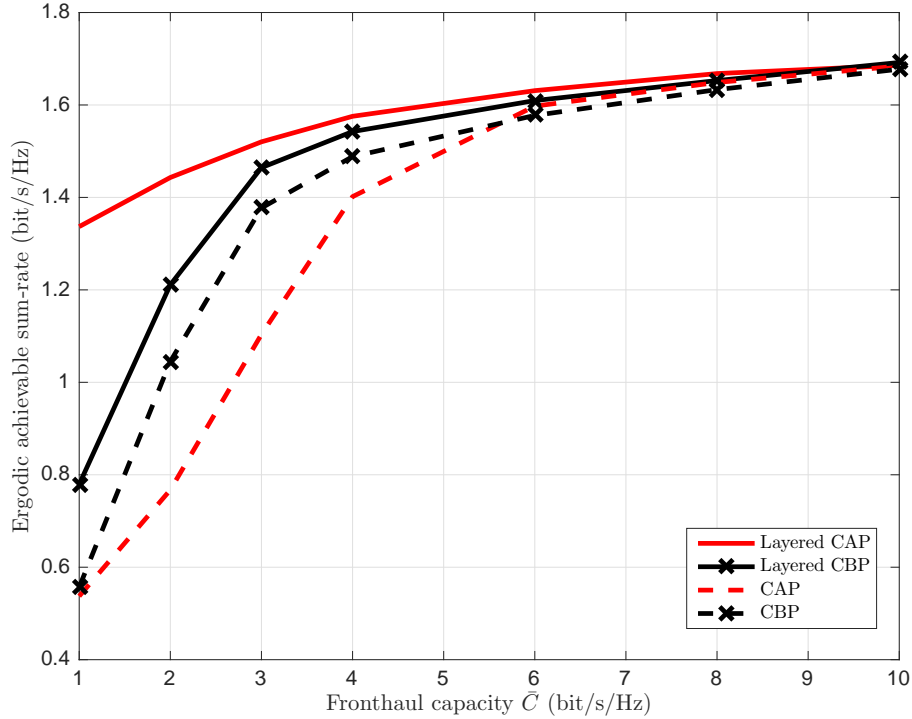


Fig. 11. Ergodic achievable sum-rate vs. the fronthaul capacity \bar{C} ($N_R = 2$, $N_M = 2$, $N_{A,i} = 2$, $N_{E,i} = 4$, $P = 5$ dB, and $T = 10$).

This is due to the separate compression of the azimuth precoded signals of layered CAP, which entails a fronthaul overhead proportional to the number of MSs.

In Fig. 11, the ergodic achievable sum-rate is plotted versus the fronthaul capacity \bar{C} for $N_R = N_M = 2$, $N_{A,i} = 2$ and $N_{E,i} = 4$ for all $i \in \mathcal{N}_R$, $\bar{P} = 5$ dB, and $T = 10$. We first remark that the performance gain of the layered strategies is observed at low-to-moderate fronthaul capacities, while, for large fronthaul capacities, the performance of the conventional strategies approach that of the layered strategies. As a general rule, the conventional CAP strategy is uniformly better than conventional CBP as long as the fronthaul capacity is sufficiently large, due to the enhanced interference mitigation capabilities of CAP [15]. Instead, the layered CAP strategy is advantageous here across all values of fronthaul capacity.

The effect of the coherence time T is investigated in Fig. 12, with $N_R = N_M = 2$, $N_{A,i} = 2$ and $N_{E,i} = 4$ for all $i \in \mathcal{N}_R$, $\bar{C} = 4$ bit/s/Hz, and $\bar{P} = 5$ dB. The CBP schemes benefit from a larger coherence time T , since the fronthaul overhead required to transmit precoding information gets amortized over a larger period. In contrast, such overhead in layered CAP and CAP schemes scales proportionally

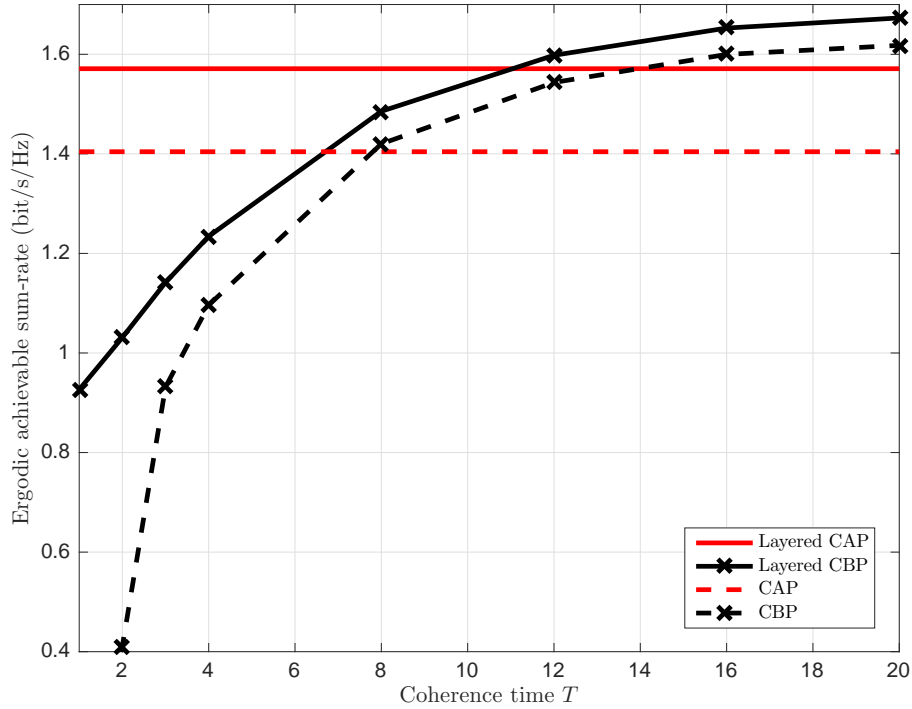


Fig. 12. Ergodic achievable sum-rate vs. the coherence time T ($N_R = N_M = 2$, $N_{A,i} = 2$, $N_{E,i} = 4$, $C = 4$ bit/s/Hz, and $P = 5$ dB).

to the coherence time T and hence the layered CAP and CAP schemes are not affected by the coherence time.

VI. CONCLUDING REMARKS

In this paper, we have studied the design of downlink Cloud Radio Access Network (C-RAN) systems in which the Radio Units (RUs) are equipped with Full Dimensional (FD)-MIMO arrays. We proposed to leverage the special low-rank structure of FD-MIMO channel, which exhibits different rates of variability in the elevation and azimuth components, by means of a novel layered precoding strategy coupled with an adaptive fronthaul compression scheme. Specifically, in the layered strategy, a single precoding matrix is optimized for the elevation channel across all coherence times based on long-term Channel State Information (CSI), while azimuth precoding matrices are optimized across independent coherence interval by adapting to instantaneous CSI. This proposed layered approach has the unique advantage in a C-RAN of potentially reducing the fronthaul overhead, due to the opportunity to amortize the overhead related to the elevation channel component across multiple coherence times. Via numerical results, it is shown that

the layered strategies significantly outperform standard non-layered schemes, especially in the regime of low fronthaul capacity and large number of vertical antennas.

We have also considered two different functional splits for both layered and non-layered precoding, namely the conventional C-RAN implementation, also known as Compress-After-Precoding (CAP) scheme, and an alternative split, referred to as Compress-Before-Precoding (CBP), whereby channel coding and precoding are performed at the RUs. Layered precoding is seen to work better under a CAP implementation when the coherence interval is not too large and the number of vertical antennas is sufficiently large; whereas the CBP approach benefits from a longer coherence interval due to its capability to amortize the fronthaul overhead for transfer of azimuth precoding information. Interesting open issues include the investigation of a scenario with multiple interfering clusters of RUs controlled by distinct Central Units (CUs) (see [28]), and the analysis of the performance in the presence of more general FD-MIMO channel models (see, e.g., [11]).

APPENDIX A

OPTIMIZATION ALGORITHM FOR THE LAYERED CAP STRATEGY

In this Appendix, we detail the derivation of Algorithm 1 for the optimization of the layered CAP strategy. We first discuss the optimization problem for the short-term variables, namely the covariance matrix $\mathbf{V}^A(\mathbf{H})$ for azimuth precoding and the quantization noise variance $\sigma_x^2(\mathbf{H})$, which are adapted to the channel realization \mathbf{H} , for given the elevation covariance matrix \mathbf{V}^E . We then consider the optimization of the long-term variable, namely the covariance matrix \mathbf{V}^E for elevation precoding, with the given covariance matrices $\mathbf{V}^A(\mathbf{H})$ for azimuth precoding and quantization noise vectors $\sigma_x^2(\mathbf{H})$.

After obtaining the elevation covariance matrix \mathbf{V}^{E*} , using the approach in Algorithm 1, the precoding matrix \mathbf{W}^{E*} for the elevation channel is calculated via the principal eigenvector approximation [29] of the obtained solution \mathbf{V}^{E*} as $\mathbf{w}_{ji}^{E*} = \nu_{\max}(\mathbf{V}_{ji}^{E*})$ for all $j \in \mathcal{N}_M$ and $i \in \mathcal{N}_R$. In a similar fashion, the algorithm obtains the precoding matrix $\mathbf{W}^{A*}(\mathbf{H})$ for the azimuth channel via the standard rank-reduction approach [29] from the obtained solution $\mathbf{V}^A(\mathbf{H})^*$ as $\mathbf{w}_{ji}^{A*}(\mathbf{H}) = \beta_{ji} \nu_{\max}(\mathbf{V}_{ji}^A(\mathbf{H})^*)$ with the normalization factors

β_{ji} selected to satisfy the power constraint with equality, namely $P_i(\mathbf{W}_i^{A*}(\mathbf{H}), \mathbf{W}_i^{E*}, \boldsymbol{\sigma}_{x,i}^{2*}(\mathbf{H})) = \bar{P}_i$.

A. Optimization over $\mathbf{V}^A(\mathbf{H})$ and $\boldsymbol{\sigma}_x^2(\mathbf{H})$ with given \mathbf{V}^E

Here, we tackle the problem (11) based on the DC algorithm [25] given the elevation precoding covariance matrix \mathbf{V}^E over the azimuth covariance matrix $\mathbf{V}^A(\mathbf{H})$ and the quantization noise variance $\boldsymbol{\sigma}_x^2(\mathbf{H})$. To this end, the objective function $R_j(\mathbf{H}, \mathbf{W}^A(\mathbf{H}), \mathbf{W}^E, \boldsymbol{\sigma}_x^2(\mathbf{H}))$ is approximated by a locally tight lower bound $\tilde{R}_j(\mathbf{H}, \mathbf{V}^A(\mathbf{H}), \boldsymbol{\sigma}_x^2(\mathbf{H}) | \mathbf{V}^{A(l-1)}(\mathbf{H}), \boldsymbol{\sigma}_x^{2(l-1)}(\mathbf{H}), \mathbf{V}^E)$ around solutions $\mathbf{V}^{A(l-1)}(\mathbf{H})$ and $\boldsymbol{\sigma}_x^{2(l-1)}(\mathbf{H})$ obtained at $(l-1)$ -th inner iteration with

$$\begin{aligned} \tilde{R}_j(\mathbf{H}, \mathbf{V}^A(\mathbf{H}), \boldsymbol{\sigma}_x^2(\mathbf{H}) | \mathbf{V}^{A(l-1)}(\mathbf{H}), \boldsymbol{\sigma}_x^{2(l-1)}(\mathbf{H}), \mathbf{V}^E) = & \log \left(1 + \sum_{k=1}^{N_M} \sum_{i=1}^{N_R} \rho_{ji}(\mathbf{V}_{ki}^A(\mathbf{H}), \mathbf{V}_{ki}^E, \sigma_{x,ki}^2(\mathbf{H})) \right) \\ & - f \left(1 + \sum_{k=1, k \neq j}^{N_M} \sum_{i=1}^{N_R} \rho_{ji}(\mathbf{V}_{ki}^{A(l-1)}(\mathbf{H}), \mathbf{V}_{ki}^E, \sigma_{x,ki}^{2(l-1)}(\mathbf{H})), 1 + \sum_{k=1, k \neq j}^{N_M} \sum_{i=1}^{N_R} \rho_{ji}(\mathbf{V}_{ki}^A(\mathbf{H}), \mathbf{V}_{ki}^E, \sigma_{x,ki}^2(\mathbf{H})) \right) \end{aligned} \quad (18)$$

where $\rho_{ji}(\mathbf{V}_{ki}^A, \mathbf{V}_{ki}^E, \sigma_{x,ki}^2) = \lambda_{ji}^E \mathbf{u}_{ji}^E \mathbf{V}_{ki}^E \mathbf{u}_{ji}^\dagger \left(\mathbf{h}_{ji}^A \mathbf{V}_{ki}^A \mathbf{h}_{ji}^{A\dagger} + \sigma_{x,ki}^2 \|\mathbf{h}_{ji}^A\|^2 \right)$ and the linearized function $f(a, b)$ is obtained from the first-order Taylor expansion of the log function as $f(a, b) = \log(a) + (b - a)/a$. Since the fronthaul constraint (11b) is a DC constraint, the left-hand side of the constraint (11b) is approximated by applying successive locally tight convex lower bounds as

$$\begin{aligned} \tilde{C}_{x,i}(\mathbf{V}_i^A(\mathbf{H}), \boldsymbol{\sigma}_{x,i}^2(\mathbf{H}) | \mathbf{V}_i^{A(l-1)}(\mathbf{H}), \boldsymbol{\sigma}_{x,i}^{2(l-1)}(\mathbf{H})) \triangleq \\ \sum_{j=1}^{N_M} \left\{ f \left(\text{tr}(\mathbf{V}_{ji}^{A(l-1)}(\mathbf{H})) + \sigma_{x,ji}^{2(l-1)}(\mathbf{H}), \text{tr}(\mathbf{V}_{ji}^A(\mathbf{H})) + \sigma_{x,ji}^2(\mathbf{H}) \right) - \log \sigma_{x,ji}^2 \right\}. \end{aligned} \quad (19)$$

At l -th inner loop, the following convex optimization problem, for given $\mathbf{V}^{A(l-1)}(\mathbf{H})$, $\boldsymbol{\sigma}_x^{2(l-1)}(\mathbf{H})$, and \mathbf{V}^E , is solved for obtaining new iterates $\mathbf{V}^{A(l)}(\mathbf{H})$ and $\boldsymbol{\sigma}_x^{2(l)}(\mathbf{H})$ as

$$\mathbf{V}^{A(l)}(\mathbf{H}), \boldsymbol{\sigma}_x^{2(l)}(\mathbf{H}) \leftarrow \arg \max_{\mathbf{V}^A(\mathbf{H}), \boldsymbol{\sigma}_x^2(\mathbf{H})} \sum_{j \in \mathcal{N}_M} \tilde{R}_j(\mathbf{H}, \mathbf{V}^A(\mathbf{H}), \boldsymbol{\sigma}_x^2(\mathbf{H}) | \mathbf{V}^{A(l-1)}(\mathbf{H}), \boldsymbol{\sigma}_x^{2(l-1)}(\mathbf{H}), \mathbf{V}^E) \quad (20a)$$

$$\text{s.t.} \quad \tilde{C}_{x,i}(\mathbf{V}_i^A(\mathbf{H}), \boldsymbol{\sigma}_{x,i}^2(\mathbf{H}) | \mathbf{V}_i^{A(l-1)}(\mathbf{H}), \boldsymbol{\sigma}_{x,i}^{2(l-1)}(\mathbf{H})) \leq \bar{C}_i, \quad (20b)$$

$$P_i(\mathbf{V}_i^A(\mathbf{H}), \mathbf{V}_i^E, \boldsymbol{\sigma}_{x,i}^2(\mathbf{H})) \leq \bar{P}_i, \quad \forall i \in \mathcal{N}_R. \quad (20c)$$

The DC method obtains the solutions $\mathbf{V}^A(\mathbf{H})$ and $\boldsymbol{\sigma}_x^2(\mathbf{H})$ by solving the problem (20) iteratively over l until a convergence criterion is satisfied and the resulting algorithm is summarized in Algorithm 2.

B. Optimization over \mathbf{V}^E

In this part, the covariance matrix \mathbf{V}^E for elevation precoding is designed for given azimuth precoding covariance matrices $\mathbf{V}^{A(m)} = \mathbf{V}^{A(m)}(\mathbf{H}^{(m)})$ and quantization noise vectors $\boldsymbol{\sigma}_x^{2(m)} = \boldsymbol{\sigma}_x^{2(m)}(\mathbf{H}^{(m)})$ for all $m = 1, \dots, n$. Since the elevation covariance matrix $\mathbf{V}^{E(n)}$ is not adapted to the channel realization \mathbf{H} and the objective function (11) is non-convex with respect to $\mathbf{V}^{E(n)}$, in this optimization, we use the SSUM algorithm [24]. To this end, at each step, a stochastic lower bound of the objective function is maximized around the current iterate. Following the SSUM method, at n -th outer loop, the objective function with given $\mathbf{V}^{A(m)}$ and $\boldsymbol{\sigma}_x^{2(m)}$, for all $m = 1, \dots, n$, is reformulated as the empirical average

$$\frac{1}{n} \sum_{m=1}^n \tilde{R}_j(\mathbf{H}^{(m)}, \mathbf{V}^E | \mathbf{V}^{E(m-1)}, \mathbf{V}^{A(m)}, \boldsymbol{\sigma}_x^{2(m)}), \quad (21)$$

where $\tilde{R}_j(\mathbf{H}^{(m)}, \mathbf{V}^E | \mathbf{V}^{E(m-1)}, \mathbf{V}^{A(m)}, \boldsymbol{\sigma}_x^{2(m)})$ is a locally tight convex lower bound around the previous iterate $\mathbf{V}^{E(m-1)}$, when the channel realization is $\mathbf{H}^{(m)}$, and is calculated as

$$\begin{aligned} \tilde{R}_j(\mathbf{H}^{(m)}, \mathbf{V}^E | \mathbf{V}^{E(m-1)}, \mathbf{V}^{A(m)}, \boldsymbol{\sigma}_x^{2(m)}) &= \log \left(1 + \sum_{k=1}^{N_M} \sum_{i=1}^{N_R} \rho_{ji}(\mathbf{H}^{(m)}, \mathbf{V}_{ki}^{A(m)}, \mathbf{V}_{ki}^E, \sigma_{x,i}^{2(m)}) \right) \\ &- f \left(1 + \sum_{k=1, k \neq j}^{N_M} \sum_{i=1}^{N_R} \rho_{ki}(\mathbf{H}^{(m)}, \mathbf{V}_{ki}^{A(m)}, \mathbf{V}_{ki}^{E(m-1)}, \sigma_{x,i}^{2(m)}), 1 + \sum_{k=1, k \neq j}^{N_M} \sum_{i=1}^{N_R} \rho_{ki}(\mathbf{H}^{(m)}, \mathbf{V}_{ki}^{A(m)}, \mathbf{V}_{ki}^E, \sigma_{x,i}^{2(m)}) \right), \end{aligned} \quad (22)$$

with $\rho_{ji}(\mathbf{H}^{(m)}, \mathbf{V}_{ki}^A, \mathbf{V}_{ki}^E, \sigma_{x,i}^2) = \lambda_{ji}^{E(m)} \mathbf{u}_{ji}^{E(m)} \mathbf{V}_{ki}^E \mathbf{u}_{ji}^{(m)\dagger} (\mathbf{h}_{ji}^{A(m)} \mathbf{V}_{ki}^A \mathbf{h}_{ji}^{A(m)\dagger} + \sigma_{x,i}^2 \|\mathbf{h}_{ji}^{A(m)}\|^2)$. The n -th iterate $\mathbf{V}^{E(n)}$ is obtained by solving the following convex optimization problem

$$\mathbf{V}^{E(n)} \leftarrow \arg \max_{\mathbf{V}^E} \quad \frac{1}{n} \sum_{m=1}^n \sum_{j \in \mathcal{N}_M} \tilde{R}_j(\mathbf{H}^{(m)}, \mathbf{V}^E | \mathbf{V}^{E(m-1)}, \mathbf{V}^{A(m)}, \boldsymbol{\sigma}_x^{2(m)}) \quad (23a)$$

$$\text{s.t.} \quad C_{x,i}(\mathbf{V}_i^{A(n)}, \boldsymbol{\sigma}_{x,i}^{2(n)}) \leq \bar{C}_i, \quad \forall i \in \mathcal{N}_R, \quad (23b)$$

$$P_i(\mathbf{V}_i^{A(n)}, \mathbf{V}_i^E, \boldsymbol{\sigma}_{x,i}^{2(n)}) \leq \bar{P}_i, \quad \forall i \in \mathcal{N}_R. \quad (23c)$$

As in Section A-A, the outer loop in Algorithm 1 is repeated until the convergence is achieved.

APPENDIX B

OPTIMIZATION ALGORITHM FOR LAYERED CBP STRATEGY

In this Appendix, the precoding matrices \mathbf{W}^{E*} and $\widetilde{\mathbf{W}}^{A*}$, MSs' rates $\{R_j\}$ and quantization noise vector $\boldsymbol{\sigma}_w^{2*}$ are jointly optimized for the CBP-based strategy. The optimization of short-term variables, namely the

covariance matrix $\tilde{\mathbf{V}}^A(\mathbf{H})$ for azimuth precoding and the quantization noise variance $\sigma_w^2(\mathbf{H})$, which are adapted to the channel realization \mathbf{H} for given the elevation covariance matrix \mathbf{V}^E , is described first. Then, the optimization over the long-term variables, namely the covariance matrix \mathbf{V}^E for elevation precoding and the user rates $\{R_j\}$, is discussed given covariance matrices $\mathbf{V}^{A(m)}(\mathbf{H})$ for azimuth precoding and quantization noise vectors $\sigma_w^{2(m)}(\mathbf{H})$, for all $m = 1, \dots, n$, as detailed in Appendix B-B.

As in Appendix A, the elevation precoding matrix \mathbf{W}^{E*} and the azimuth precoding matrix $\tilde{\mathbf{W}}^{A*}$ are calculated via the standard rank-reduction approach [29] with the obtained solutions \mathbf{V}^{E*} and $\tilde{\mathbf{V}}^{A*}$, respectively, as detailed in Algorithm 3.

A. Optimization over $\tilde{\mathbf{V}}^A(\mathbf{H})$ and $\sigma_w^2(\mathbf{H})$ with given \mathbf{V}^E

Here, we aim at maximizing the objective function (17a) over the azimuth precoding covariance matrix $\tilde{\mathbf{V}}^A(\mathbf{H})$ and the quantization noise variance $\sigma_w^2(\mathbf{H})$ given the elevation precoding covariance matrix \mathbf{V}^E using the DC method [25]. At the l -th iteration of the DC method, the non-convex functions $\bar{R}_j(\mathbf{H}, \tilde{\mathbf{V}}^A(\mathbf{H}), \mathbf{V}^E, \sigma_w^2(\mathbf{H}))$ and $C_{w,i}(\tilde{\mathbf{V}}_i^A(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H}))$ are respectively substituted with a locally tight lower bound $\tilde{R}_j(\mathbf{H}, \tilde{\mathbf{V}}^A(\mathbf{H}), \sigma_w^2(\mathbf{H}) | \tilde{\mathbf{V}}^{A(l-1)}(\mathbf{H}), \sigma_w^{2(l-1)}(\mathbf{H}), \mathbf{V}^E)$ and a tight upper bound $\tilde{C}_{w,i}(\tilde{\mathbf{V}}_i^A(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H}) | \tilde{\mathbf{V}}_i^{A(l-1)}(\mathbf{H}), \sigma_{w,i}^{2(l-1)}(\mathbf{H}))$, obtained as in Appendix A. The bounds are given by

$$\begin{aligned} \tilde{R}_j(\mathbf{H}, \tilde{\mathbf{V}}^A(\mathbf{H}), \sigma_w^2(\mathbf{H}) | \tilde{\mathbf{V}}^{A(l-1)}(\mathbf{H}), \sigma_w^{2(l-1)}(\mathbf{H}), \mathbf{V}^E) = & \log \left(1 + \sum_{i=1}^{N_R} \sum_{k \in \mathcal{M}_i} \rho_{ji}(\tilde{\mathbf{V}}_{ki}^A(\mathbf{H}), \mathbf{V}_{ki}^E, \sigma_{w,i}^2(\mathbf{H})) \right) \\ & - f \left(1 + \sum_{i=1}^{N_R} \sum_{k \in \mathcal{M}_i \setminus j} \rho_{ki}(\tilde{\mathbf{V}}_{ki}^{A(l-1)}(\mathbf{H}), \mathbf{V}_{ki}^E, \sigma_{w,i}^{2(l-1)}(\mathbf{H})), 1 + \sum_{i=1}^{N_R} \sum_{k \in \mathcal{M}_i \setminus j} \rho_{ki}(\tilde{\mathbf{V}}_{ki}^A(\mathbf{H}), \mathbf{V}_{ki}^E, \sigma_{w,i}^2(\mathbf{H})) \right), \end{aligned} \quad (24)$$

and

$$\begin{aligned} \tilde{C}_{w,i}(\tilde{\mathbf{V}}_i^A(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H}) | \tilde{\mathbf{V}}_i^{A(l-1)}(\mathbf{H}), \sigma_{w,i}^{2(l-1)}(\mathbf{H})) \triangleq \\ \frac{1}{T} \left\{ f \left(\tilde{\mathbf{V}}_i^{A(l-1)}(\mathbf{H}) + \sigma_{w,i}^{2(l-1)}(\mathbf{H}) \mathbf{I}, \tilde{\mathbf{V}}_i^A(\mathbf{H}) + \sigma_{w,i}^2(\mathbf{H}) \mathbf{I} \right) - N_{A,i} \log(\sigma_{w,i}^2) \right\}, \end{aligned} \quad (25)$$

where $\rho_{ji}(\tilde{\mathbf{V}}_{ki}^A, \mathbf{V}_{ki}^E, \sigma_{w,i}^2) = \lambda_{ji}^E \mathbf{u}_{ji}^E \mathbf{V}_{ki}^E \mathbf{u}_{ji}^\dagger \left(\mathbf{h}_{ji}^A \tilde{\mathbf{V}}_{ki}^A \mathbf{h}_{ji}^{A\dagger} + \sigma_{w,i}^2 \|\mathbf{h}_{ji}^A\|^2 \right)$ and the linearization function $f(\mathbf{A}, \mathbf{B})$

for the matrices is defined as $f(\mathbf{A}, \mathbf{B}) \triangleq \log \det(\mathbf{A}) + \text{tr}(\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A}))$.

At l -th iteration of DC method, the following convex optimization problem for given $\tilde{\mathbf{V}}^{A(l-1)}(\mathbf{H})$, $\sigma_w^{2(l-1)}(\mathbf{H})$ and \mathbf{V}^E is solved for obtaining new iterates $\tilde{\mathbf{V}}^{A(l)}(\mathbf{H})$ and $\sigma_w^{2(l)}(\mathbf{H})$:

$$\tilde{\mathbf{V}}^{A(l)}(\mathbf{H}), \sigma_w^{2(l)}(\mathbf{H}) \leftarrow \arg \max_{\tilde{\mathbf{V}}^A(\mathbf{H}), \sigma_w^2(\mathbf{H}), \{R_j\}} \sum_{j \in \mathcal{N}_M} R_j \quad (26a)$$

$$\text{s.t.} \quad R_j \leq \tilde{R}_j(\mathbf{H}, \tilde{\mathbf{V}}^A(\mathbf{H}), \sigma_w^2(\mathbf{H}) | \tilde{\mathbf{V}}^{A(l-1)}(\mathbf{H}), \sigma_w^{2(l-1)}(\mathbf{H}), \mathbf{V}^E), \quad \forall j \in \mathcal{N}_M, \quad (26b)$$

$$\tilde{C}_{w,i}(\tilde{\mathbf{V}}_i^A(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H}) | \tilde{\mathbf{V}}_i^{A(l-1)}(\mathbf{H}), \sigma_{w,i}^{2(l-1)}(\mathbf{H})) \leq \bar{C}_i - \sum_{j \in \mathcal{M}_i} R_j, \quad (26c)$$

$$P_i(\tilde{\mathbf{V}}_i^A(\mathbf{H}), \mathbf{V}_i^E, \sigma_{w,i}^2(\mathbf{H})) \leq \bar{P}_i, \quad \forall i \in \mathcal{N}_R. \quad (26d)$$

Problem (26) is solved iteratively over l until convergence and the resulting algorithm is summarized in Algorithm 4.

B. Optimization over \mathbf{V}^E and $\{R_j\}$

We design the covariance matrix \mathbf{V}^E for elevation precoding and the user rates $\{R_j\}$ for given azimuth precoding covariance matrices $\tilde{\mathbf{V}}^{A(m)} = \tilde{\mathbf{V}}^{A(m)}(\mathbf{H}^{(m)})$ and quantization noise vectors $\sigma_w^{2(m)} = \sigma_w^{2(m)}(\mathbf{H}^{(m)})$ for all $m = 1, \dots, n$. As in Appendix A, this optimization problem can be tackled via the SSUM method. To this end, the function $E[\tilde{R}_j(\mathbf{H}, \tilde{\mathbf{W}}^A(\mathbf{H}), \mathbf{W}^E, \sigma_w^2(\mathbf{H}))]$ in (17b) is approximated with the stochastic upper bound as

$$\frac{1}{n} \sum_{m=1}^n \tilde{R}_j(\mathbf{H}^{(m)}, \mathbf{V}^E | \mathbf{V}^{E(m-1)}, \tilde{\mathbf{V}}^{A(m)}, \sigma_w^{2(m)}), \quad (27)$$

with

$$\begin{aligned} \tilde{R}_j(\mathbf{H}^{(m)}, \mathbf{V}^E | \mathbf{V}^{E(m-1)}, \tilde{\mathbf{V}}^{A(m)}, \sigma_w^{2(m)}) &= \log \left(1 + \sum_{i=1}^{N_R} \sum_{k \in \mathcal{M}_i} \rho_{ji}(\mathbf{H}^{(m)}, \tilde{\mathbf{V}}_{ki}^{A(m)}, \mathbf{V}_{ki}^E, \sigma_{w,i}^{2(m)}) \right) \\ &- f \left(1 + \sum_{i=1}^{N_R} \sum_{k \in \mathcal{M}_i \setminus j} \rho_{ki}(\mathbf{H}^{(m)}, \tilde{\mathbf{V}}_{ki}^{A(m)}, \mathbf{V}_{ki}^E, \sigma_{w,i}^{2(m)}), 1 + \sum_{i=1}^{N_R} \sum_{k \in \mathcal{M}_i \setminus j} \rho_{ki}(\mathbf{H}^{(m)}, \tilde{\mathbf{V}}_{ki}^{A(m)}, \mathbf{V}_{ki}^E, \sigma_{w,i}^{2(m)}) \right), \end{aligned} \quad (28)$$

where $\rho_{ji}(\mathbf{H}^{(m)}, \tilde{\mathbf{V}}_{ki}^A, \mathbf{V}_{ki}^E, \sigma_{w,i}^2) = \lambda_{ji}^{E(m)} \mathbf{u}_{ji}^{E(m)} \mathbf{V}_{ki}^E \mathbf{u}_{ji}^{(m)\dagger} (\mathbf{h}_{ji}^{A(m)} \tilde{\mathbf{V}}_{ki}^A \mathbf{h}_{ji}^{A(m)\dagger} + \sigma_{w,i}^2 ||\mathbf{h}_{ji}^{A(m)}||^2)$. At the n -th iteration, $\mathbf{V}^{E(n)}$ and $\{R_j^{(n)}\}$ are obtained by solving the following optimization problem based on SSUM

method

$$\mathbf{V}^{E(n)}, \{R_j^{(n)}\} \leftarrow \arg \max_{\mathbf{V}^E, \{R_j\}} \sum_{j \in \mathcal{N}_M} R_j \quad (29a)$$

$$\text{s.t.} \quad R_j \leq \frac{1}{n} \sum_{m=1}^n \tilde{R}_j(\mathbf{H}^{(m)}, \mathbf{V}^E | \mathbf{V}^{E(m-1)}, \tilde{\mathbf{V}}^{A(m)}, \boldsymbol{\sigma}_w^{2(m)}), \quad \forall j \in \mathcal{N}_M, \quad (29b)$$

$$C_{w,i}(\tilde{\mathbf{V}}_i^A(\mathbf{H}), \sigma_{w,i}^2(\mathbf{H})) \leq \bar{C}_i - \sum_{j \in \mathcal{M}_i} R_j, \quad (29c)$$

$$P_i(\tilde{\mathbf{V}}_i^A(\mathbf{H}), \mathbf{V}_i^E, \sigma_{w,i}^2(\mathbf{H})) \leq \bar{P}_i, \quad \forall i \in \mathcal{N}_R \quad (29d)$$

until convergence.

REFERENCES

- [1] China Mobile, “C-RAN: the road towards green RAN,” White Paper, ver. 2.5, China mobile Research Institute, Oct. 2011.
- [2] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, “Cloud RAN for mobile networks - a technology overview,” *IEEE Communications Surveys and Tutorials*, vol. 17, no. 1, pp. 405–426, First quarter 2015.
- [3] D. Samardzija, J. Pastalan, M. MacDonald, S. Walker, and R. Valenzuela, “Compressed transport of baseband signals in radio access networks,” *IEEE Trans. Wireless Comm.*, vol. 11, no. 9, pp. 3216–3225, Sep. 2012.
- [4] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, “Fronthaul compression for cloud radio access networks: signal processing advances inspired by network information theory,” *IEEE Sig. Proc. Mag.*, vol. 31, no. 6, pp. 69–79, Nov. 2014.
- [5] U. Dotsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier, “Quantitative analysis of split base station processing and determination of advantageous architectures for LTE,” *Bell Labs Technical Journal*, vol. 18, no. 1, pp. 105–128, Jun. 2013.
- [6] D. Wubben, P. Rost, J. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, “Benefits and impact of cloud computing on 5G signal processing: Flexible centralization through cloud-RAN,” *IEEE Sig. Proc. Mag.*, vol. 31, no. 6, pp. 35–44, Nov. 2014.
- [7] J. Bartelt, P. Rost, D. Wubben, J. Lessmann, B. Melis, and G. Fettweis, “Fronthaul and backhaul requirements of flexibly centralized radio access networks,” *IEEE Wireless Comm.*, vol. 22, no. 5, pp. 105–111, Oct. 2015.
- [8] A. D. L. Oliva, X. C. Perez, A. Azcorra, A. D. Giglio, F. Cavaliere, D. Tiegelbekkers, J. Lessmann, T. Haustein, A. Mourad, and P. Iovanna, “Xhaul: toward an integrated fronthaul/backhaul architecture in 5g networks,” *IEEE Wireless Comm.*, vol. 22, no. 5, pp. 32–40, Oct. 2015.
- [9] P.-H. Kuo, “New physical layer features of 3GPP LTE release-13 [Industry Perspectives],” *IEEE Wireless Comm.*, vol. 22, no. 4, pp. 4–5, Aug. 2015.
- [10] Y.-H. Nam, B. L. Ng, K. Sayana, Y. Li, J. Zhang, Y. Kim, and J. Lee, “Full-dimension MIMO (FD-MIMO) for next generation cellular technology,” *IEEE Comm. Mag.*, vol. 51, no. 4, pp. 172–179, Jun. 2013.
- [11] G. Xu, Y. Li, Y.-H. Nam, C. Zhang, T. Kim, and J.-Y. Seol, “Full-dimension MIMO: Status and challenges in design and implementation,” in *2014 IEEE Communication Theory Workshop (CTW)*, the Piscadera Bay, Curaçao, May 2014.
- [12] O. Simeone, O. Somekh, H. V. Poor, and S. Shamai, “Downlink multicell processing with limited-backhaul capacity,” *EURASIP Jour. Adv. Sig. Proc.*, Jun. 2009.
- [13] P. Marsch and G. Fettweis, “On downlink network MIMO under a constrained backhaul and imperfect channel knowledge,” *Proc. IEEE Glob. Comm. Conf.*, pp. 1–6, Honolulu, HI, USA, Nov. 2009.
- [14] P. Patil and W. Yu, “Hybrid compression and message-sharing strategy for the downlink cloud radio-access network,” *Proc. of IEEE Info. Th. and Application Workshop*, pp. 1–6, San Diego, CA, USA, Feb. 2014.
- [15] J. Kang, O. Simeone, J. Kang, and S. Shamai, “Fronthaul compression and precoding design for C-RANs over ergodic fading channel,” to appear in *IEEE Trans. Veh. Techn.*, 2015.

- [16] D. Ying, F. W. Vook, T. A. Thomas, D. J. Love, and A. Ghosh, "Kronecker product correlation model and limited feedback codebook design in a 3D channel model," *Proc. IEEE Int. Conf. on Comm.*, pp. 5865–5870, Sydney, NSW, Australia, Jun. 2014.
- [17] A. Alkhateeb, G. Leus, and R. W. H. Jr., "Multi-layer precoding for full-dimensional massive MIMO systems," *Proc. of Asilomar Conf. on Sign., Syst. and Computers*, pp. 815–819, Pacific Grove, CA, USA, Nov. 2014.
- [18] S. Park, C.-B. Chae, and S. Bahk, "Before/after precoded massive MIMO in cloud radio access networks," *Proc. IEEE Int. Conf. on Comm.*, pp. 169–173, Budapest, Hungary, Jun. 2013.
- [19] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.
- [20] N. Seifi, J. Zhang, R. W. H. Jr., T. Svensson, and M. Coldrey, "Coordinated 3D beamforming for interference management in cellular networks," *IEEE Trans. Wireless Comm.*, vol. 13, no. 10, pp. 5396–5410, Oct. 2014.
- [21] Z. Zhong, X. Yin, X. Li, and X. Li, "Extension of ITU IMT-advanced channel models for elevation domains and line-of-sight scenarios," *Proc. IEEE Veh. Technol. Conf.*, pp. 1–5, Las Vegas, NV, USA, Sep. 2013.
- [22] T. M. Cover and J. A. Thomas, *Element of Information Theory*. John Wiley & Sons, 2006.
- [23] M. Brookes, "The matrix reference manual," [online] <http://www.ee.imperial.ac.uk/hp/staff/dmb/matrix/intro.html>, 2011.
- [24] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, "A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks," *arXiv:1307.4457*.
- [25] A. Beck and M. Teboulle, "Gradient-based algorithms with applications to signal recovery problems," in *Convex Optimization in Signal Processing and Communications*, Y. Eldar and D. Palomar, editors, pp. 42–48, Cambridge University Press 2010.
- [26] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, Feb. 2004.
- [27] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [28] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Inter-cluster design of precoding and fronthaul compression for cloud radio access networks," *IEEE Wireless Comm. Lett.*, vol. 3, no. 4, pp. 369–372, Aug. 2014.
- [29] L. Vandenberghe and S. Boyd, "Semidefinite relaxation of quadratic optimization problems," *SIAM Rev.*, vol. 38, no. 1, pp. 49–95, 1996.